

Estimation of errors in iterative solutions of a non-symmetric linear system

Aashish Vishwakarma*, Murugesan Venkatapathi⁺
 Supercomputer Education and Research Centre
 Indian Institute of Science, Bangalore, India
 aashish.vishwakarma@gmail.com*, muruges@serc.iisc.ernet.in⁺

Abstract

Estimation of actual errors from the residue in iterative solutions is necessary for efficient solution of large problems when their condition number is much larger than one. Such estimators for conjugate gradient algorithms used to solve symmetric positive definite linear systems exist. This work presents error estimation for iterative solutions of general indefinite linear systems to provide accurate stopping and restarting criteria. In many realistic applications no properties of the matrices are known a priori; thus requiring such a general algorithm.

Our method for approximating the required quadratic form $r^T A^{-1} r$ (square of the A -norm of the error vector) when solving nonsymmetric linear systems with Bi-Conjugate Gradient (BiCG) algorithm, needs only $O(1)$ time (per BiCG iteration). We also extend this estimate to approximate l_2 -norm of error vector using the relations of Hestenes and Stiefel [1]. Using the heuristics of numerical results we observe that the developed algorithm (BiCGQL) is at least $\kappa \times 10^{-1}$ times more accurate than residue vector based stopping criteria (where κ is the condition number of the system).

Keywords: Conjugate Gradients; BiCG; error-bounds; stopping criteria; condition number

1 Introduction

When using any iterative algorithm for solving a linear system $[Ax = b]$, one of the most important questions is when to stop the iterations. One would like to stop the iterations when the norm of the error (where x_k are the iterates)

$$\epsilon_k = x - x_k \tag{1}$$

is small enough. However, the actual error is unknown and most iterative algorithms rely on residual vector as stopping criteria like $\|r_k\|_2 \approx \|\epsilon\|_2 \|b\|_2$ where $r_k = b - Ax_k$ is the residual vector. Such stopping criteria can work when the system is well-conditioned. However, these types of stopping criteria can be misleading depending on the condition number of A or the choice of the initial approximation. It

can stop the iterations too early when the norm of the error is still much larger than tolerance, or too late in which case too many floating point operations have been done for obtaining the required accuracy.

Also, in the case when the condition number of the matrix is too large, the residual vector might not show proper converging behavior. In fact the residual vector might show oscillating behavior while the actual error might still be (however slowly) converging (and vice-versa). In such cases, the residual vector cannot be used as a good stopping or restarting criteria. The norm of relative residue can be as large as κ times or as small as $1/\kappa$ times the norm of the relative error.

Even though most iterative algorithms are used with preconditioner, in most realistic situations, it is not guaranteed whether preconditioner will actually reduce the condition number of the matrix (this can mostly be seen in case of large matrices). This created motivation for ways to compute estimates of some norms of the error for iterative solvers. Such estimators (e.g. CGQL) are already available for CG algorithm ([2]). For solving nonsymmetric linear systems using FOM and GMRES methods formulas for estimation of errors have been suggested [3] recently. Our objective is to derive efficient estimators for solving general nonsymmetric linear systems.

In our paper, we briefly recall the error estimates existing for Hermitian Positive Definite (HPD) problems (section -2). We show the equivalence conditions between (non symmetric) Lanczos co-efficients and BiCG iterates (sections 3.2-3.3). We develop efficient $O(1)$ estimations for A -norm and l_2 -norm of the error vector for general indefinite problems (sections 3.4-3.5) using local bi-orthogonality conditions. The estimation formulas we derive depend only upon BiCG iterates and add no extra cost to the BiCG algorithm. We test this method (BiCGQL) for BiCG computations and compare them with the residual based stopping criteria and existing bounds for non-symmetric problems suggested by Golub and Meurant in their book ([4], p.210). We show that our estimators result in large improvements of the stopping criteria as the condition number of the problems increase.

2 Related Work

Algorithm 1 Lanczos Algorithm

Input A, v
 $\beta_0 = 0, v_0 = 0$
 $v_1 = v/||v||$
for $k = 1 \dots$ **convergence**
 $w = Av_k - \beta_{k-1}v_{k-1}$
 $\alpha_k = v_k^T w$
 $w = w - \alpha_k v_k$
 $\beta_k = ||w||$
 $v_{k+1} = w/\beta_k$
end for

One of the most commonly used methods for solving linear systems with real symmetric positive definite (SPD) matrix is Conjugate Gradient (CG) algorithm. It can be derived from several different perspectives, (i) an orthogonalization problem, (ii) minimization problem and (iii) Lanczos algorithm. In their paper, Golub and Meurant [5] have suggested computing bounds for A-norm of the error in the Conjugate Gradient (CG) method. A typical norm of error for CG is the A-norm (also called the energy norm) which is minimized at each iteration. It is defined as

$$\|\epsilon_k\|_A^2 \equiv (\epsilon_k^T A \epsilon_k) = r^T A^{-1} A A^{-1} r = r^T A^{-1} r \quad (2)$$

It is sometimes also interesting to compute l_2 -norm, for which $\|\epsilon\|_2^2 = r^T A^{-2} r$. In order to obtain bound on $\|\epsilon\|_A$ we must obtain bound on $r^T A^{-1} r$. Of course we do not want to compute A^{-1} . So our problem is similar to obtaining computable bound for quadratic forms $u^T A^{-1} u$.

In [2], Meurant has showed how one can obtain approximation for A-norm of error in CG iterations.

When $A = A^T$,

$$\begin{aligned} \|\epsilon^2\|_A &= r^T A^{-1} r \\ &= r^T Q \Lambda^{-1} Q^T r \\ &= q^T \Lambda^{-1} q \\ &= \sum_{i=1}^n \lambda_i^{-1} q_i^2 \end{aligned} \quad (3)$$

$$= \int_{\lambda_{\min}}^{\lambda_{\max}} \lambda^{-1} d\alpha(\lambda) \quad (4)$$

$$= \int_a^b f(\lambda) d\alpha(\lambda) \text{ (In general)} \quad (5)$$

Equation 4 is Riemann–Stieltjes integral of equation 3. Here α is a piecewise constant and defined as

$$\alpha(\lambda) = 0 \quad \text{if } \lambda \leq \lambda_{\min} \quad (6)$$

$$= \sum_{j=1}^i q_j^2 \quad \text{if } \lambda_i \leq \lambda < \lambda_{i+1} \quad (7)$$

$$= \sum_{j=1}^n q_j^2 \quad \text{if } \lambda \geq \lambda_{\max} \quad (8)$$

This allows us to use Gauss, Gauss-Radau, and Gauss-Lobatto formulas for a function f given by (from equation 5)

$$\int_a^b f(\lambda) d\alpha(\lambda) = \sum_{i=1}^N w_i f(t_i) + \sum_{j=1}^M v_j f(z_j) + R[f] \quad (9)$$

where the weights w_i, v_j and nodes t_i are unknowns and nodes z_j are given. $R[f]$ can be given by

$$R[f] = \frac{f(\eta)^{2N+M}}{(2N+M)!} \int_a^b \prod_{j=1}^M (\lambda - z_j) \left(\prod_{i=1}^N (\lambda - t_i) \right)^2 d\alpha(\lambda) \quad (10)$$

where, $a < \eta < b$

When $M = 0$, the approximation of the integral (equation 9) is called the Gauss formula, when $M = 1$, $z_1 = \lambda_{min}$ or $z_1 = \lambda_{max}$ it is called Gauss-Radau and when $M = 2$, $z_1 = \lambda_{min}$ and $z_2 = \lambda_{max}$ it is called Gauss-Lobatto. The nodes t and z can be obtained by a polynomial decomposition of the integral in terms of $p_i(\lambda)$. Moreover, a set of orthogonal polynomials provides a 3-term recursion relationship for easy calculations. This means the recurrence coefficients can be represented in a matrix of symmetric tri-diagonal form; the crucial observation being that these can be trivially extracted from the CG iterates, resulting in negligible addition of computation over the iterative solution. In more generality, the CG algorithm can be described as a minimization of the polynomial relation

$$\|x - x_k\|_A = \min_{p_k} \|p_k(A)(x - x_0)\|_A \quad (11)$$

Given $\int_a^b p_i(\lambda) p_j d\alpha(\lambda) = 0$ when $i \neq j$ and 1 when $i = j$ and $\gamma_i p_i(\lambda) = (\lambda - \omega_i) p_{i-1}(\lambda) + \gamma_{i-1} p_{i-2}(\lambda)$, when $i = 1 \dots N$, normalized such that

$$\int d\alpha = 1, p_0(\lambda) = 1, p_{-1}(\lambda) = 0$$

$$\Rightarrow \lambda P_{N-1}(\lambda) = T_N P_{N-1}(\lambda) + \gamma_N p_N(\lambda) \varepsilon_N \quad (12)$$

where $\varepsilon_N^T = [0, 0, \dots, 1]$, $P_{N-1}(\lambda)^T = [p_1(\lambda), \dots, p_{N-1}(\lambda)]$ and T_N is the Jacobi matrix obtained by Lanczos Algorithm as discussed later. These techniques are used for providing lower and upper bounds for quadratic forms $u^T f(A) u$ where f is a smooth function, A is an SPD matrix and u is a given vector. Paper by Golub, Gene H., and Zdeněk Strakoš [6] talks about how to obtain error estimations in quadratic formulas. The algorithm GQL (Gauss Quadrature and Lanczos) is based on the Lanczos algorithm and on computing functions of Jacobi matrices. These techniques are adapted to the CG algorithm to compute lower and upper bounds on the A-norm of the error. The idea is to use CG instead of the Lanczos algorithm, to compute explicitly the entries of the corresponding Jacobi matrices from the CG coefficients, and then to use the same formulas as in GQL. The formulas are summarized in the CGQL algorithm (QL standing for Quadrature and Lanczos) (algorithm 3). The CGQL algorithm uses the tridiagonal Jacobi matrix obtained by translating the coefficients computed in CG into the Lanczos coefficients. This paper focuses on establishing the relationships between (non-symmetric) Lanczos co-efficients and BiCG iterates in order to obtain expressions for approximation of A-norm of the error vector and further extending the approach for obtaining approximation of l_2 -norm of the error vector.

The Lanczos, CG and CGQL Algorithms

Given a starting vector v and an SPD matrix A , the Lanczos algorithm (algorithm 1) computes an orthonormal basis v_1, \dots, v_{k+1} of the Krylov subspace $\kappa_{k+1}(A, v)$, which is defined as...

$$\kappa_{k+1}(A, v) = \text{span}\{v, Av, \dots, A^k v\} \quad (13)$$

In Algorithm 1 we have used the modified Gram-Schmidt form of the algorithm. The basis vectors v_j satisfy the matrix relation

$$AV_k = V_k T_k + \eta_{k+1} v_{k+1} \varepsilon_k^T \quad (14)$$

Here, ε_k is the k^{th} canonical vector, where $V_k = [v_1 \dots v_k]$ and T_k is the $k \times k$ symmetric tridiagonal matrix of the recurrence coefficients computed in algorithm 1:

$$T_k = \begin{bmatrix} \alpha_1 & \eta_1 & & & \\ \eta_1 & \ddots & & & \\ & & \ddots & \eta_{k-1} & \\ & & & \eta_{k-1} & \alpha_k \end{bmatrix} \quad (15)$$

The coefficients β_j being positive, T_k is a Jacobi matrix. The Lanczos algorithm works for any symmetric matrix, but if A is positive definite, then T_k is positive definite as well.

When solving a system of linear algebraic equations $Ax = b$ with symmetric and positive definite matrix A , the CG method (algorithm 2) can be used. CG (which may be derived from the Lanczos algorithm) computes iterates x_k that are optimal since the A-norm of the error defined in (1) is minimized over $x_0 + \kappa_k(A, r_0)$,

$$\|x - x_k\|_A = \min_{y \in x_0 + \kappa_k(A, r_0)} \|x - y\|_A \quad (16)$$

Algorithm 2 Conjugate Gradient Algorithm

input A, b, x_0

$r_0 = b - Ax_0$

$p_0 = r_0$

for $k = 1 \dots n$ **do**

$$\gamma_{k-1} = \frac{r_{k-1}^T r_{k-1}}{p_{k-1}^T A p_{k-1}}$$

$$x_k = x_{k-1} + \gamma_{k-1} p_{k-1}$$

$$r = r_{k-1} - \gamma_{k-1} A p_{k-1}$$

$$\beta_k = \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}}$$

$$p_k = r_k + \beta_k p_{k-1}$$

end for

It is well-known that the recurrence coefficients computed in both algorithms (Lanczos and CG) are connected via

$$\eta_k = \frac{\sqrt{\beta_k}}{\gamma_{k-1}}, \alpha_k = \frac{1}{\gamma_{k-1}} + \frac{\beta_{k-1}}{\gamma_{k-2}}, \delta_0 = 0, \gamma_{-1} = 1 \quad (17)$$

Noticing that the error ϵ_k and the residual r_k are related through $A\epsilon_k = r_k$, we have

$$\|\epsilon\|_A^2 = \epsilon_k^T A \epsilon_k = r_k^T A^{-1} r_k \quad (18)$$

The above formula has been used for reconstructing the A-norm of the error. For the sake of simplicity, only Gauss rule has been considered. Let s_k be the estimate of $\|\epsilon^k\|_A$. Let d be a positive integer, then the idea is to use the following formula at CG iteration k ,

In their paper [5], Golub and Meurant give following expression pertaining to A-norm of the error.

$$\|\epsilon\|_A^2 = \|r_0\|^2 ((T_n^{-1})_{1,1} - (T_k^{-1})_{1,1}) \quad (19)$$

Further, for sufficiently large k , and $d = 1$, they denote the estimator as

$$s_{k-1} = \|r^0\|_2^2 \frac{\eta_{k-1}^2 c_{k-1}}{\delta_{k-1}(\alpha_k \delta_{k-1} - \eta^2)} > 0 \quad (20)$$

Using rules of Gauss, Gauss-Radalu, and Gauss Lobatto error bound, the CGQL algorithm can be established as algorithm 3. It should be noted that, in their more recent work, Gerard Muerant[7] has derived formula relating the l_2 -norm of the error in CG algorithm.

Algorithm 3 CGQL (Conjugate Gradients and Quadrature via Lanczos coefficients)

input $A, b, x_0, \lambda_m, \lambda_M$

$r_0 = b - Ax_0, p_0 = r_0$

$\eta_0 = 0, \gamma_{-1} = 1, c_1 = 1, \beta_0 = 0, \delta_0 = 1, \bar{\alpha}(\mu)^1 = \lambda_m, \underline{\alpha}(\eta)^1 = \lambda_M$

for $k = 1 \dots$ until convergence **do**

CG-iteration (k)

$$\alpha_k = \frac{1}{\gamma_{k-1}} + \frac{\beta_{k-1}}{\gamma_{k-2}},$$

$$\eta_k^2 = \frac{\beta_k}{\gamma_{k-1}^2}$$

$$\delta_k = \alpha_k - \frac{\beta_{k-1}^2}{\delta_{k-1}},$$

$$g_k = \|r_0\| \frac{c_k^2}{\delta_k}$$

$$\bar{\delta}_k = \alpha_k - \bar{\alpha}_k, \bar{\alpha}_{k+1} = \lambda_m + \frac{\beta^2}{\bar{\delta}_k},$$

$$\bar{f}_k = \|r_0\|^2 \frac{\eta_k^2 c_k^2}{\delta_k (\bar{\alpha}_{k+1} \delta_k - \eta_k^2)}$$

$$\underline{\delta}_k = \alpha_k - \underline{\alpha}_k, \underline{\alpha}_{k+1} = \lambda_M + \frac{\beta^2}{\underline{\delta}_k},$$

$$\underline{f}_k = \|r_0\|^2 \frac{\eta_k^2 c_k^2}{\delta_k (\underline{\alpha}_{k+1} \delta_k - \eta_k^2)}$$

$$\check{\alpha}_{k+1} = \frac{\bar{\delta}_k \underline{\delta}_k}{\bar{\delta}_k - \underline{\delta}_k} \left(\frac{\lambda_m}{\bar{\delta}_k} - \frac{\lambda_M}{\underline{\delta}_k} \right),$$

$$\check{\eta}_k = \frac{\bar{\delta}_k \underline{\delta}_k}{\bar{\delta}_k - \underline{\delta}_k} (\lambda_M - \lambda_m)$$

$$\bar{f}_k = \|r_0\|^2 \frac{[\check{\eta}_k]^2 c_k^2}{\delta_k (\check{\alpha}_{k+1} \delta_k - [\check{\eta}_k]^2)}$$

$$c_{k+1}^2 = \frac{\eta_k^2 c_k^2}{\delta_k^2}$$

end for

3 Methodology

3.1 The Problem

One of the most useful algorithm for iterative solution of non-symmetric linear systems in context of Lanczos and CG algorithms is Bi-Conjugate Gradient (Bi-CG) algorithm.

A-norm of error in BiCG can be written as,

$$\|\epsilon_A\|^2 = e^T A e = r^T A^{-1} r \quad (21)$$

Here, r is residual vector pertaining to the BiCG method. When A is positive definite, the right side of the above equation is always positive, it is also called as energy norm in physics related problems. In case of indefinite matrices, the absolute value of the above equation is considered. Moreover, the l_2 -norm of the error can be written as,

$$\|\epsilon\|^2 = e^T e = r^T A^{-2} r \quad (22)$$

We are interested in approximating 21 and 22. In their paper, Starkov and Tichy [8] develop a method of $O(\sim n)$ to approximate a bilinear form $(c^T A b)$ based on BiCG method. Our goal is to approximate the quantity $r_k^T A^{-1} r_k$ (A-norm of the error)

for every iteration of a BiCG algorithm. In following sections, we will derive the approximation for A -norm and l_2 -norm of the error for every BiCG iteration. The BiCG method is shown as algorithm 4.

Algorithm 4 BiCG Algorithm

input: A, A^T, x_0, b
 $r_0 = b - Ax_0, \tilde{r}_0 = p_0 = \tilde{p}_0 = r;$
for $k = 1, \dots$
 $\alpha_k = \frac{\tilde{r}_k^T r_k}{p_k^T A p_k}$
 $x_{k+1} = x_k + \alpha_k p_k, \quad \tilde{x}_{k+1} = \tilde{x}_k + \alpha_k \tilde{p}_k$
 $r_{k+1} = r_k - \alpha_k A p_k, \quad \tilde{r}_{k+1} = \tilde{r}_k - \alpha_k A^T \tilde{p}_k$
 $\beta_{k+1} = \frac{\tilde{r}_{k+1}^T r_{k+1}}{\tilde{r}_k^T r_k}$
 $p_{k+1} = r_{k+1} + \beta_{k+1} p_k, \quad \tilde{p}_{k+1} = \tilde{r}_{k+1} + \beta_{k+1} \tilde{r}_k$
end

3.2 Non-Symmetric Lanczos algorithm

Let A be a non-singular matrix of order n . We introduce the Lanczos algorithm as a means of computing an orthogonal basis of a Krylov subspace. Let v_1 and \tilde{v}_1 be given vectors (such that $\|v_1\| = 1$ and $(v_1, \tilde{v}_1) = 1$).

For $k = 1, 2, \dots$

$$\begin{aligned} z_k &= Av_k - \omega_k v_k - \eta_{k-1} v_{k-1} \\ w_k &= A^T \tilde{v}_k - \omega_k \tilde{v}_k - \tilde{\eta}_{k-1} \tilde{v}_{k-1} \end{aligned} \quad (23)$$

The coefficient ω_k being computed as $\omega_k = (\tilde{v}_k, Av_k)$. The other coefficients η_k and $\tilde{\eta}_k$ are chosen (provided $(z_k, w_k) = 0$) such that $\eta_k \tilde{\eta}_k = (z_k, w_k)$ and the new vectors at step $k + 1$ are given by

$$\begin{aligned} v_{k+1} &= \frac{z_k}{\eta_k} \\ \tilde{v}_{k+1} &= \frac{w_k}{\tilde{\eta}_k} \end{aligned} \quad (24)$$

These relations can be written in matrix form, let

$$T_k = \begin{pmatrix} \omega_1 & \eta_1 & & & \\ \tilde{\eta}_1 & \omega_2 & \eta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \tilde{\eta}_{k-1} & \omega_{k-1} & \eta_{k-1} \\ & & & \tilde{\eta}_k & \omega_k \end{pmatrix} \quad (25)$$

and

$$\begin{aligned} V_k &= [v_1 \dots v_k] \\ \tilde{V}_k &= [\tilde{v}_1 \dots \tilde{v}_k] \end{aligned} \quad (26)$$

then

$$\begin{aligned} AV_k &= V_k T_k + \tilde{\eta}_k v_{k+1} (\varepsilon_k)^T \\ A^T \tilde{V}_k &= \tilde{V}_k T_k^T + \eta_k \tilde{v}_{k+1} (\varepsilon_k)^T \end{aligned} \quad (27)$$

In order to approximate A^{-1} , which is restricted onto $\kappa_n(A, r_0)$, following holds for non-symmetric Lanczos algorithm

$$A^{-1} = V_n T_n^{-1} \tilde{V}_n^T \quad (28)$$

Considering the starting vectors $v_1 = r_0 / \|r_0\|$ and $w_1 = \|r_0\| / r_0$, we get

$$\begin{aligned} r_0^T A^{-1} r_0 &= \frac{r_0^T r_0}{\|r_0\|} w_1 V_n T_n^{-1} \tilde{V}_n v_1 \|r_0\| \\ &= (r_0^T r_0) \varepsilon_1^T T_n^{-1} \varepsilon_1 \end{aligned} \quad (29)$$

$$= \|r_0\|^2 (T_n^{-1})_{(1,1)} \quad (30)$$

where ε_1 is first canonical vector. In the next section we are going to establish relationship between BiCG iterates and Lanczos co-efficients.

3.3 BiCG , Gauss Quadrature and Lanczos (BiCGQL)

For a square matrix A , having the distribution function $w(\lambda)$ and interval (a, b) such that $a < \lambda_1 < \lambda_2 \dots < \lambda_n < b$, for any continuous function, one can define Riemann-Stieltjes integral as

$$\int_a^b f(\lambda) dw(\lambda) \quad (31)$$

where $w(\lambda)$ is a stepwise constant function.

$$w(\lambda) = \begin{cases} 0 & \text{for } \lambda < \lambda_1 \\ \sum_{j=1}^i w_j & \text{for } \lambda_i \leq \lambda < \lambda_{i+1}, 1 \leq i \leq n-1 \\ \sum_{j=1}^n w_j & \text{for } \lambda_n > \lambda \end{cases}$$

Integral 31 is a finite sum,

$$\int_a^b f(\lambda) dw(\lambda) = \sum_{i=1}^n w_i f(\lambda_i) = v_1^T f(A) v_1$$

We are interested in the quadratic formula, $r_k^T A^{-1} r_k$. It can be written using Riemann-Stieltjes integral for function $f(\lambda) = 1/\lambda$. In n^{th} step of non-symmetric Lanczos algorithm we get the full orthonormal basis of $\kappa_n(A, v_1)$ and we have

$$AV_n = V_n T_n \quad (32)$$

$$\Rightarrow A^{-1}V_n = V_n T_n^{-1} \quad (33)$$

and

$$\begin{aligned} \int_a^b f(\lambda)dw(\lambda) &= \sum_{i=1}^n w_i f(\lambda_i) = v_1^T f(A)v_1 \\ &= v_1^T A^{-1}\varepsilon_1 = v_1^T V_n(T_n^{-1})\varepsilon_1 \\ &= \varepsilon_1^T (T_n^{-1})\varepsilon_1 = (T_n^{-1})_{1,1} \end{aligned} \quad (34)$$

From above equation and equation 30, it can be said that BiCG can implicitly compute weights and nodes of Gauss Quadrature rule applied to Riemann-Stieltjes integral as

$$\int_a^b f(\lambda)dw(\lambda) = (T_n^{-1})_{1,1} = \frac{\|x - x_0\|_A^2}{\|r_0\|^2} \quad (35)$$

As mentioned earlier, using Gauss rule on the interval $[a, b]$ and a function f (such that its Riemann-Stieltjes integral and all moments exist), the above function can be approximated as

$$\int_a^b f(\lambda)dw(\lambda) = \sum_{i=1}^k w_i f(v_i) + R_k^G \quad (36)$$

In Lanczos terms, it can be expressed as

$$(T_n^{-1})_{1,1} = (T_k^{-1})_{1,1} + R_k^G \quad (37)$$

The remainder is nothing but scaled $A - norm$ of the error.

$$R_k^G = \frac{r_k^T A^{-1}r_k}{\|r_0\|^2} \quad (38)$$

Using BiCG iterates from algorithm, the relation between r_0 and r_k can be written as

$$r_0^T A^{-1}r_0 = \sum_{j=0}^k \alpha_j \|r_j\|^2 + r_k^T A^{-1}r_k \quad (39)$$

for which, the Gauss Quadrature approximation is (using 30, 37, 38 and 39)

$$(T_k^{-1})_{1,1} = \frac{1}{\|r_0\|^2} (r_0^T A^{-1}r_0 - r_k^T A^{-1}r_k) = \frac{1}{\|r_0\|^2} \sum_{j=0}^{k-1} \alpha_j \|r_j\|^2 \quad (40)$$

3.4 $O(1)$ expression for approximating A -norm of the error

Let us again consider Gauss Quadrature rule at step k .

$$(T_n^{-1})_{1,1} = (T_k^{-1})_{1,1} + \frac{r_k^T A^{-1} r_k}{\|r_0\|^2} \quad (41)$$

Here, we want to approximate $r_k^T A^{-1} r_k$. Of course at iteration $k < n$, $(T_n^{-1})_{1,1}$, is not known. Re-writing the above equation at step $k+1$,

$$(T_n^{-1})_{1,1} = (T_{k+1}^{-1})_{1,1} + \frac{r_{k+1}^T A^{-1} r_{k+1}}{\|r_0\|^2} \quad (42)$$

Subtracting 42 from 41, we get

$$\begin{aligned} r_k^T A^{-1} r_k - r_{k+1}^T A^{-1} r_{k+1} &= \|r_0\|^2 [(T_{k+1}^{-1})_{1,1} - (T_k^{-1})_{1,1}] \\ &= [\alpha_k \|r_k\|^2] \text{ (from equation 40)} \end{aligned} \quad (43)$$

43 gives insights for approximating $r_k^T A^{-1} r_k$. Alternatively, the same expression can be derived using the expressions for r_{k+1} and p_k , in BiCG algorithm as following for $k = 0, 1, 2, 3 \dots n-1$

$$\begin{aligned} r_k^T A^{-1} r_k - r_{k+1}^T A^{-1} r_{k+1} &= (r_{k+1} + \alpha_k A^T p_k)^T A^{-1} (r_{k+1} + \alpha_k A p_k) - r_{k+1}^T A^{-1} r_{k+1} \\ &= (r_{k+1}^T + \alpha_k p_k^T A) (A^{-1} r_{k+1} + \alpha_k p_k) - r_{k+1}^T A^{-1} r_{k+1} \\ &= r_{k+1}^T A^{-1} r_{k+1} + \alpha_k p_k^T r_{k+1} + \alpha_k r_{k+1}^T p_k \\ &\quad + \alpha_k^2 p_k^T A p_k - r_{k+1}^T A^{-1} r_{k+1} \\ &\Rightarrow r_k^T A^{-1} r_k - r_{k+1}^T A^{-1} r_{k+1} = \alpha_k r_k^T r_k \end{aligned} \quad (44)$$

($\because r_{k+1}^T p_k = p_k^T r_{k+1} = 0$)

From equations 44,

$$\begin{aligned} r_k^T A^{-1} r_k - r_{k+1}^T A^{-1} r_{k+1} &= \alpha_k \|r_k\|^2 \\ \Rightarrow \|\epsilon_k\|_A^2 - \|\epsilon_{k+1}\|_A^2 &= \alpha_k \|r_k\|^2 \end{aligned} \quad (45)$$

(as $r^T A^{-1} r = \|\epsilon\|_A^2$ (A -norm of the error))

For a finite natural number $d(\geq 0)$ the above expression can be approximated as

$$\|\epsilon_{k-d}\|_A^2 \approx \sum_{j=k-d}^k \alpha_j \|r_j\|^2 \quad (46)$$

Here, d signifies the delay in approximation. It should be noted that when A is positive-definite, the above expression is always positive and thus provides a lower bound for the square of A -norm of the error. When A is indefinite, the above expression can be negative (upper bound) or positive (lower bound) depending up on residual

vector. Also, the BiCG method might show irregular convergence, in such cases higher values of d can result in less accurate approximations. Hence, lower values of d are recommended. Method like BiCGSTAB can repair the irregular convergence behavior of BiCG, as a result smoother convergence is obtained, and hence higher values of d can be used for better approximations.

3.5 $O(1)$ expression for approximation of l_2 -norm of the error

Hestenes and Stiefel [1] proved the following result relating the l_2 -norm and A -norm of the error.

$$\|\epsilon_k\|_A^2 + \|\epsilon_{k+1}\|_A^2 = [\|\epsilon_k\|^2 - \|\epsilon_{k+1}\|^2][\mu(p_k)] \quad (47)$$

$$\text{where, } \mu(p_k) = \frac{(p_k^T A p_k)}{\|p_k\|^2} \quad (48)$$

Summing equations 45 and 47,

$$\begin{aligned} 2\|\epsilon_k\|_A^2 &= [\|\epsilon_k\|^2 - \|\epsilon_{k+1}\|^2][\mu(p_k)] + \alpha_k \|r_k\|^2 \\ \Rightarrow \|\epsilon_k\|^2 - \|\epsilon_{k+1}\|^2 &= \frac{2\|\epsilon_k\|_A^2}{[\mu(p_k)] + \alpha_k \|r_k\|^2} \end{aligned} \quad (49)$$

$$\Rightarrow \|\epsilon_k\|^2 - \|\epsilon_{k+1}\|^2 = \phi_k, \quad \left(\text{where, } \phi_k = \frac{2\|\epsilon_k\|_A^2}{[\mu(p_k)] + \alpha_k \|r_k\|^2} \right) \quad (50)$$

For a finite natural number $d(\geq 0)$, above expression can be used to approximate l_2 -norm of the error as following

$$\|\epsilon_{k-d}\|^2 \approx \sum_{j=k-d}^k \phi_j \quad (51)$$

Here, d signifies the delay in approximation. It should be noted that if d_1 delay is introduced in the estimation of A -norm of error in equation 46, and d_2 delay is introduced in estimation of l_2 -norm, the total delay becomes $d_1 + d_2$ and the equation 51 becomes

$$\|\epsilon_{k-d_1-d_2}\|^2 \approx \sum_{j=k-d_1-d_2}^{k-d_1} \phi_j$$

These estimators are incorporated with BiCG in algorithm 5 (we call it BiCGQL-BiCG Quadrature Lanczos).

3.6 BiCG Convergence

Few theoretical results are known about the convergence of BiCG. For HPD systems the method delivers the same results as CG, but at twice the cost per iteration. For

Algorithm 5 BiCGQL Algorithm

input: A, A^T, x_0, b, d_1, d_2 $r_0 = b - Ax_0, \tilde{r}_0 = p_0 = \tilde{p}_0 = r;$ **for** $k = 0, 1, \dots$

$$\alpha_k = \frac{\tilde{r}_k^T r_k}{p_k^T A p_k}$$

$$\mu(p_k) = \frac{(p_k^T A p_k)}{\|p_k\|^2}$$

$$x_{k+1} = x_k + \alpha_k p_k,$$

$$\tilde{x}_{k+1} = \tilde{x}_k + \alpha_k \tilde{p}_k$$

$$r_{k+1} = r_k - \alpha_k A p_k,$$

$$\tilde{r}_{k+1} = \tilde{r}_k - \alpha_k A^T \tilde{p}_k$$

if $k \geq d_1 + d_2$

$$g_{k-d_1} = \sum_{j=k-d_1}^k \alpha_j \|r_j\|^2 \text{ (A-norm estimation)}$$

$$\phi_{k-d_1} = \frac{2\|\epsilon_{k-d_1}\|_A^2}{[\mu(p_{k-d_1})] + \alpha_{k-d_1} \|r_{k-d_1}\|^2}$$

$$f_{k-d_1-d_2} = \sum_{j=k-d_1-d_2}^{k-d_1} \phi_j \text{ (} l_2\text{-norm estimation)}$$

end if

$$\beta_{k+1} = \frac{\tilde{r}_{k+1}^T r_{k+1}}{\tilde{r}_k^T r_k}$$

$$p_{k+1} = r_{k+1} + \beta_{k+1} p_k, \quad \tilde{p}_{k+1} = \tilde{r}_{k+1} + \beta_{k+1} \tilde{r}_k$$

end

nonsymmetric matrices it has been shown that in phases of the process where there is significant reduction of the norm of the residual, the method is more or less comparable to full GMRES (in terms of numbers of iterations) (Freund and Nachtigal [9]). In practice this is often confirmed, but it is also observed that the convergence behavior may be quite irregular, and the method may even break down. The breakdown situation due to the possible event can be circumvented by so-called look-ahead strategies (Parlett, Taylor and Liu [10]). The other breakdown situation, occurs when the decomposition fails, and can be repaired by using another decomposition (such as QMR developed by Freund and Nachtigal [9][11]). Sometimes, breakdown or near-breakdown situations can be satisfactorily avoided by a restart at the iteration step immediately before the (near) breakdown step. BiCGSTAB is an improvement over BiCG algorithm which leads to a considerably smoother convergence behavior. It should be noted that relations 46 and 51 hold valid for BiCGSTAB algorithms.

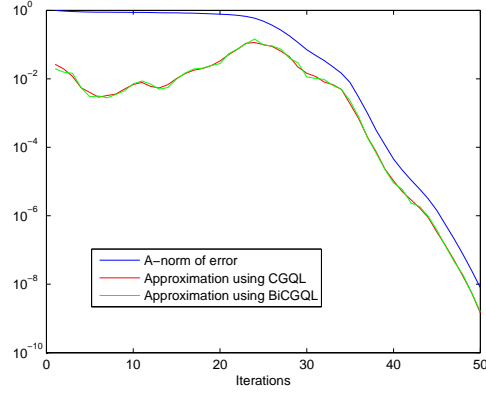


Figure 1: Comparing BiCGQL A-norm estimator with Gauss Approximation of CGQL for an HPD matrix (condition number of the matrix is 10^4 , $d_1 = d_2 = 0$)

4 Numerical Results and Validations

4.1 Tests and Results

In figure 1, the developed approximator is compared with the A -norm of the error vector as well as Gauss Approximation (discussed in CGQL algorithm). Here, A is an square HPD matrix. Figure shows that our A -norm estimator is almost as good as CGQL Gauss Rule for HPD matrices.

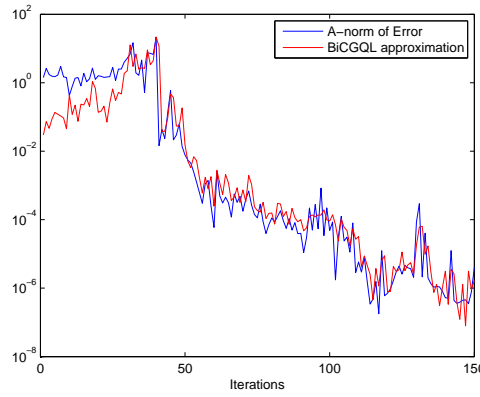


Figure 2: BiCGQL estimator in case of a non-hermitian (indefinite) matrix (absolute values are considered for quadratic term $r^T A^{-1} r$ and its approximation, condition number of matrix is 10^4 , $d_1 = d_2 = 0$)

In figure 2, A is a nonsymmetric matrix. Here, the plot of approximation vector

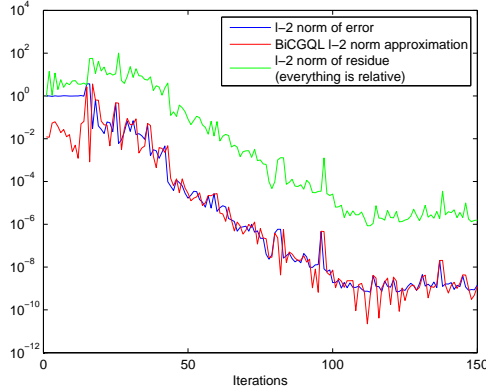


Figure 3: Comparison between BiCGQL l_2 -norm estimator, actual l_2 -norm of the error and l_2 -norm of the residue, condition number of matrix is 10^4 , $d_1 = d_2 = 0$

along with A -norm of the error is shown. Figure 3 shows the comparison between l_2 -norm approximation, actual l_2 -norm of the error and l_2 -norm of the residue. It is evident that BiCGQL estimators work efficiently both times.

For the purpose of extended tests (for both HPD and Indefinite cases), six different bins of size ten were created with varying condition number for matrix A : 1 to 10, 10 to 100, ... 10^5 to 10^6 etc. For each matrix A , 100 different instances of vector ' b ' were created, each being unique canonical form of order 100. Thus each bin represents the result accumulated from 1000 different cases. Below we are comparing our approximation of A -norm of error with estimation by residue vector. Average error in estimating A -norm of error by BiCGQL A -norm estimator can be expressed as

$$\left| \frac{\frac{g_k}{\|x\|_A} - \frac{\|e_k\|_A}{\|x\|_A}}{\frac{\|e_k\|_A}{\|x\|_A}} \right| \quad (52)$$

where $\|e\|_A$ is the A -norm of error vector, g_k is BiCGQL A -norm estimator, $\|r\|_2$ is the l_2 -norm of residue vector, $\|e\|_2$ is the l_2 -norm of error vector and x is actual solution vector. While error in estimating l_2 -norm of the error by residual can be expressed as

$$\left| \frac{\frac{\|r_k\|_2}{\|b\|_2} - \frac{\|e_k\|_2}{\|x\|_2}}{\frac{\|e_k\|_2}{\|x\|_2}} \right| \quad (53)$$

Ratio of equation 52 to equation 53 would show the performance of BiCGQL A -norm estimator compared to residual as the estimator of the l_2 -norm of the error. In 4 and 5, each bar represents the average value of "ratio of equation 52 to equation 53 averaged over all iterations" over 1000 different cases. Results obtained for 49 show that the approximation of the A -norm of the error obtained by our A -norm estimator

is much better than approximation of the l_2 -norm of the error obtained by the residue vector. It should be noted that without our approximator iterative methods would rely on the residue vector which poorly approximated l_2 -norm of the error. The figures 4 & 5 show that residue keeps becoming unreliable as the condition number of the problem increases. The graph also shows that our approximator remains effective in approximating A -norm of the error regardless of the condition number of the problem.

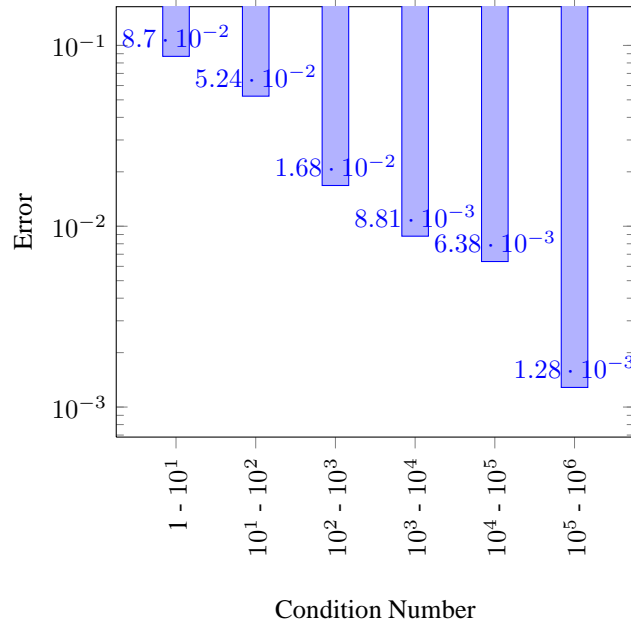


Figure 4: Average ratio of relative error in estimating A -norm by BiCGQL and relative error in traditional stopping criteria for an HPD matrix (each bar shows average over 1000 cases, $d_1 = d_2 = 4$)

Similarly, we now show [fig. 6 & 7] that our l_2 -norm approximator is much better compared to residue vector in approximating l_2 -norm of the error vector. Test conditions remain the same as in the previous case. Here, each bar represents average of following over 1000 different cases...

$$\left| \frac{\frac{f_k}{\|x\|_2} - \frac{\|e_k\|_2}{\|x\|_2}}{\frac{\|r_k\|_2}{\|b\|_2} - \frac{\|e_k\|_2}{\|x\|_2}} \right| \quad (54)$$

here f_k is BiCGQL l_2 -norm estimator. Even here, BiCGQL l_2 -norm is proven to be superior than

It is noteworthy that estimator for l_2 -norm of the error holds greater significance than the estimator for A -norm of the error in most realistic applications. We shall now compare our estimators with the estimators suggested by Golub and Meurant ([4], p.210) as below:

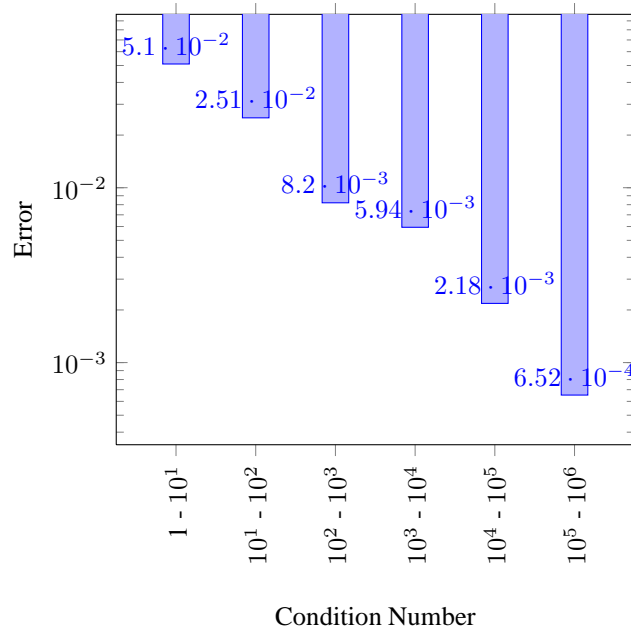


Figure 5: Average ratio of relative error in estimating A -norm by BiCGQL and relative error in traditional stopping criteria for a non-HPD matrix (each bar shows average over 1000 cases, $d_1 = d_2 = 4$)

$$\|\epsilon\|_A^2 \approx \frac{(r, Ar)}{(A^2 r, Ar)} \quad (55)$$

$$\|\epsilon\|^2 \approx \frac{(r, r)^2}{(Ar, Ar)} \quad (56)$$

The following expression is averaged over all 1000 different cases, same as above. (Here g_k^{GM} is the estimator suggested by Golub and Muerant in 55.)

$$\left| \left(\frac{\frac{q_k}{\|x\|_A} - \frac{\|e_k\|_A}{\|x\|_A}}{\frac{g_k^{GM}}{\|x\|_A} - \frac{\|e_k\|_A}{\|x\|_A}} \right) \right| \quad (57)$$

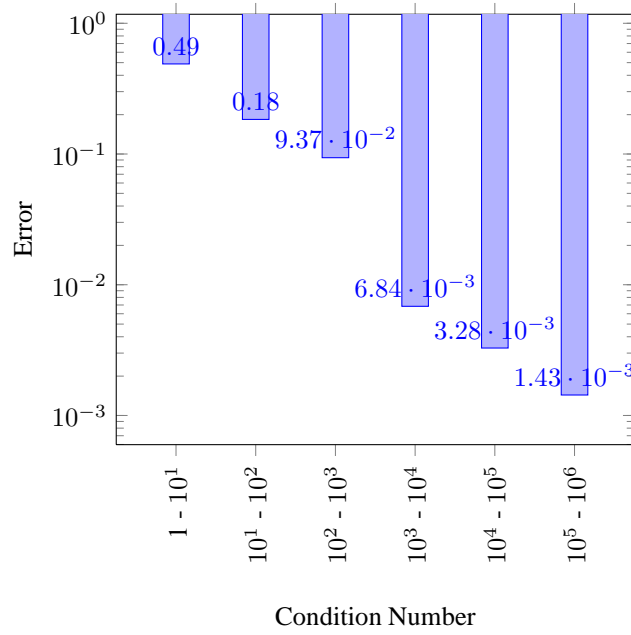


Figure 6: Average ratio of relative error in estimating l_2 -norm by BiCGQL and relative error in traditional stopping criteria for an HPD matrix (each bar shows average over 1000 cases, $d_1 = d_2 = 4$)

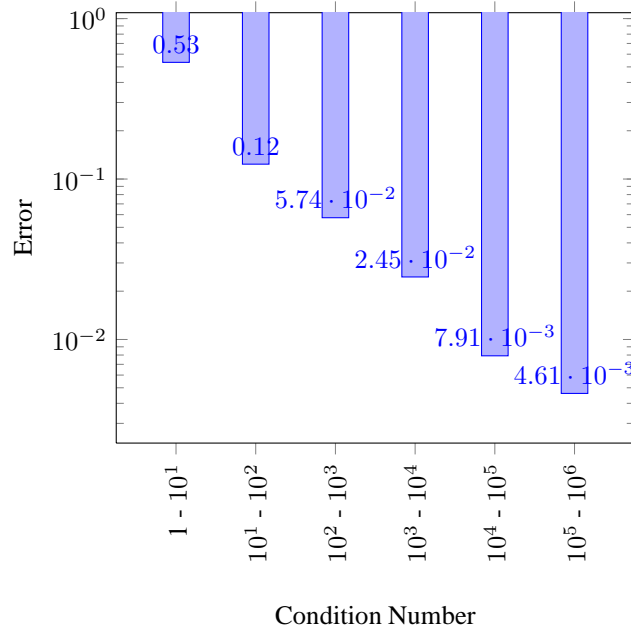


Figure 8: Average ratio of the relative error in estimating A -norm of the error by BiCGQL and Golub-Meurant estimations for a non-HPD matrix (each bar shows average over 1000 cases, $d_1 = d_2 = 0$)

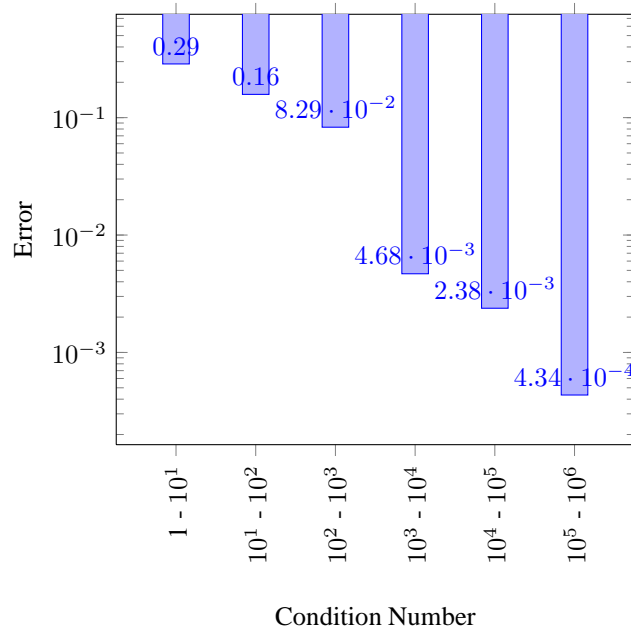


Figure 7: Average ratio of relative error in estimating l_2 -norm by BiCGQL and relative error in traditional stopping criteria for a non-HPD matrix (each bar shows average over 1000 cases, $d_1 = d_2 = 4$)

The following expression is averaged over all 1000 different cases, same as above. (Here f_k^{GM} is the estimator suggested by Golub and Muerant in 56.)

$$\left| \left(\frac{\frac{f_k}{\|x\|_2} - \frac{\|e_k\|_2}{\|x\|_2}}{\frac{f_k^{GM}}{\|x\|_2} - \frac{\|e_k\|_2}{\|x\|_2}} \right) \right| \quad (58)$$

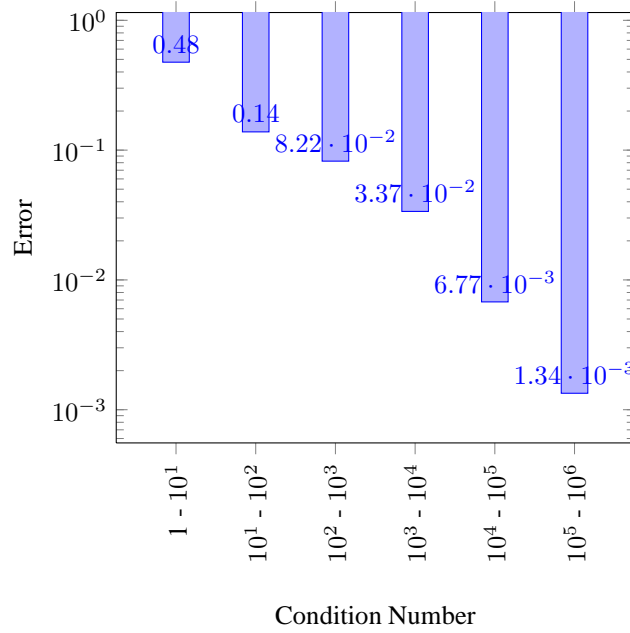


Figure 9: Average ratio of the relative error in estimating l_2 -norm of the error by BiCGQL and Golub-Meurant estimations for a non-HPD matrix(each bar shows average over 1000 cases, $d_1 = d_2 = 0$)

From figures 8 and 9, it can be clearly seen that the estimators in equations 55 and 56 are good for well-conditioned problem, but they fail for problems with high conditioned problems. It should also be noted that these estimators are computationally very expensive as they include matrix-matrix multiplications. Our estimators (BiCGQL) give much better results comparatively. Thus it is evident that BiCGQL estimators are superior in terms of both accuracy and computational cost.

5 Conclusions

The importance of BiCGQL estimators are evident for problems with moderately high condition number $\kappa > 100$, and is emphasized by a few general examples in section 4. The $O(1)$ estimators for BiCG computations developed by us are on an average $\kappa \times 10^{-1}$ times more accurate than residual based stopping criteria and $\kappa \times 10^{-2}$ times more accurate than the previously existing estimators. As matrix A is non-Hermitian, BiCGQL estimators do not necessarily give upper or lower bounds on the norms of errors, however as previously discussed, they can be used for indefinite problems. Based on the results presented in the previous section, we believe that the estimate for the A -norm and l_2 -norm of the error should be implemented into software realization of the BiCG or similar iterative algorithms as a stopping criteria instead of the residual.

References

- [1] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*. NBS, 1952, vol. 49.
- [2] G. Meurant and Z. Strakoš, “The lanczos and conjugate gradient algorithms in finite precision arithmetic,” *Acta Numerica*, vol. 15, pp. 471–542, 2006.
- [3] G. Meurant, “Estimates of the norm of the error in solving linear systems with fom and gmres,” *SIAM Journal on Scientific Computing*, vol. 33, no. 5, pp. 2686–2705, 2011.
- [4] G. H. Golub and G. Meurant, *Matrices, moments and quadrature with applications*. Princeton University Press, 2009.
- [5] —, “Matrices, moments and quadrature ii; how to compute the norm of the error in iterative methods,” *BIT Numerical Mathematics*, vol. 37, no. 3, pp. 687–705, 1997.
- [6] G. H. Golub and Z. Strakoš, “Estimates in quadratic formulas,” *Numerical Algorithms*, vol. 8, no. 2, pp. 241–268, 1994.
- [7] G. Meurant, “Estimates of the l_2 norm of the error in the conjugate gradient algorithm,” *Numerical Algorithms*, vol. 40, no. 2, pp. 157–169, 2005.
- [8] Z. Strakoš and P. Tichý, “On efficient numerical approximation of the bilinear form $c^*a^{-1}b$,” *SIAM Journal on Scientific Computing*, vol. 33, no. 2, pp. 565–587, 2011.
- [9] R. W. Freund and N. M. Nachtigal, “Qmr: a quasi-minimal residual method for non-hermitian linear systems,” *Numerische Mathematik*, vol. 60, no. 1, pp. 315–339, 1991.
- [10] B. N. Parlett, D. R. Taylor, and Z. A. Liu, “A look-ahead lanczos algorithm for unsymmetric matrices,” *Mathematics of computation*, vol. 44, no. 169, pp. 105–124, 1985.
- [11] R. W. Freund and N. M. Nachtigal, “An implementation of the qmr method based on coupled two-term recurrences,” *SIAM Journal on Scientific Computing*, vol. 15, no. 2, pp. 313–337, 1994.